A JOINT EMD AND TEAGER-KAISER ENERGY APPROACH TOWARDS NORMAL AND NASAL SPEECH ANALYSIS

Chris De La Cruz and Balu Santhanam

Department of ECE, University of New Mexico
Albuquerque, NM: 87131-001, USA.
Email: cfdelac,bsanthan@unm.edu.

ABSTRACT

Velopharyngeal inadequacy produced in cleft lip and palate (CLP) situations manifests as hypernasality in the underlying speech and utterances. Common methods for the analysis of these utterances are one-third octave band analysis and LPC analysis. The CEEMDAN-2014 algorithm is used to decompose both normal and nasal utterances into underlying intrinsic mode functions and then Teager-Kaiser energy operator-based energy metrics of the IMFs in the lower and high frequency bands are computed. The proposed energy metrics are shown to produce a clear delineation between nasal and normal resonances taken from utterances containing various levels of hypernasality in the American CLP Craniofacial database.

Keywords: Empirical mode decomposition, Teager-Kaiser energy operator, hypernasality, energy metrics, nasal and normal vowel analysis.

1. INTRODUCTION

Traditional formant analysis of normal vowels, assumes narrowband resonances and employs an LPC approach towards estimating the formant frequencies and bandwidths. Hypernasality refers to the presence of excessive nasalization when producing vowels or voiced consonants, or both [1, 2]. This condition is associated with improper velopharyngeal closure during the production of speech utterances such as those seen in CLP cases. Although perceptual judgement is the standard for rating hypernasality, it has poor reliability [1].

Recently, there is significant interest in the development of valid and consistent instrumental and acoustic measures to supplement perceptual judgement. One such technique, is the one-third-octave band analysis of the underlying utterances [1, 2]. In particular, results from [1, 2] indicate that the spectral amplitude increases at lower frequencies (630 Hz) and decreases at higher frequencies (greater than 2.5 kHz)

In this paper, we propose a hybrid approach towards the analysis of nasal utterances combining the *empirical mode decomposition* (EMD) to separate the utterance into constituent *intrinsic mode functions* (IMF's) [6] with energy metrics derived from the Teager-Kaiser energy operator (TKEO) [3]. These measure the energies in the IMFs present in the low and high frequency bands corresponding to F1, F2, and F3 [2], associated with both nasal and normal utterances. The hybrid approach is applied to signals in the CLP database with various levels of hypernasality categorized by perceptual evaluation [10]. The proposed energy metrics when applied to the samples in [10] are shown to produce a clear delineation between various levels of hypernasality such as mild, moderate, or severe [2] and the results, shown to be consistent with those seen in [1, 2].

2. PROPOSED APPROACH

The proposed approach uses a combination of ensemble EMD-based based separation of the utterances into constituent IMFs and the computation of energy metrics to quantify the energy of the IMFs in the low and high frequency bands corresponding to the frequency bands F1, F2, and F3 [2].

2.1. EEMD Analysis of Normal and Nasal Resonances

The EMD algorithm and variants employ a sifting process to decompose the input signal into constituent IMFS's $m_i[n]$, satisfying local mean and maxima-minima properties [6]: (a) $m_i[n]$ must have zero local mean and (b) $m_i[n]$ must have one zero between successive extrema.

The ensemble EMD (EEMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) approaches [5, 7] inject noise to mitigate mode mixing that occurs if input components are spectrally close [6]. The IMFs contain both amplitude and frequency modulation and are wideband waveforms. The EEMD approach and the extracted IMF's have recently been used for the analysis of normal speech vowels [8] while [9] applies LPC analysis to selected IMFs to distinguish between nasal and normal resonances.

2.2. TKEO Metrics

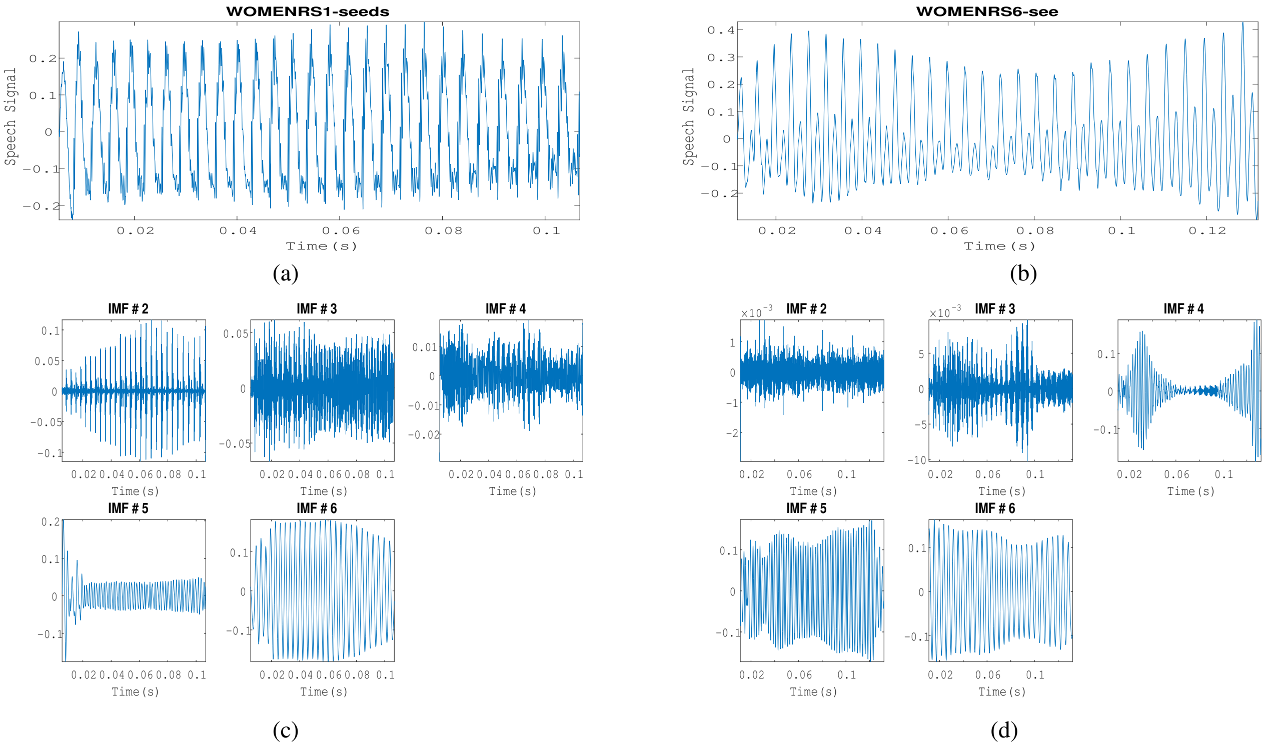
To quantify the differences between the energies of the IMF's contained in the lower and higher frequency bands analogous to F1, F2, and F3 in [2], we propose the use of the Teager-Kaiser energy operator (TKEO)-based energies [3] derived from median filtering of the TKEO output of the different IMF's. The first metric measures the energy present at higher frequencies (greater that 1 kHz, $f_s = 44.1\ kHz$):

$$\eta_1[k] = \frac{\sum_{i=1}^{k} \tilde{\psi}\left(m_i[n]\right)}{\sum_{i=1}^{n} \tilde{\psi}\left(m_i[n]\right)}, \tag{1}$$

where $\tilde{\psi}$ denotes median filtering of the TKEO given by:

$$\psi(x[n]) = x^2[n] - x[n-1]x[n+1], \tag{2}$$

The authors acknowledge Amy Neel, UNM Speech and Hearing Sciences, for speech physiology insight and Hillary Jones, American Cleft Palate-Craniofacial Association, for access to the voice samples.



(a)

(b)

(c)

(d)

Fig. 1. Voice samples from [10]. (a) Normal voice signal of utterance ”seeds” from speaker ’WOMENRS1’. (b) Nasal voice signal of utterance ”see” from speaker ’WOMENRS6’. (c) IMFs for normal voice signal. (d) IMFs for nasal voice signal. Lower-numbered IMFs retain more of the original signal’s high-frequency content while higher-numbered IMFs retain more of the low-frequency content. Only 5 (displayed) of 12 output IMFs for each signal contribute significantly to the total signal energy.

IMF	η_1:normal	η_2:normal	η_1:nasal3	η_2:nasal3	η_1:nasal6	η_2:nasal6
1	0.0069	1.0000	0.0027	1.0000	0.0010	1.0000
2	0.0695	0.9865	0.0256	0.9933	0.0017	0.9983
3	0.7294	0.8683	**0.2069**	0.9731	0.0072	0.9979
4	**0.7590**	0.2728	0.2069	**0.7123**	**0.0972**	**0.9889**
5	0.8088	**0.2449**	0.8085	0.6752	0.8302	0.8261
6	1.0000	0.2032	1.0000	0.1467	0.9999	0.1615
7	1.0000	0.0000	1.0000	0.0000	1.0000	0.0001
8	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 1. Energy metrics: normal voice ’WOMENRS1’, Level-3 nasal voice ’WOMENRS3’, and Level-6 nasal voice ’WOMENRS6’ [10]. Bolded η_1s represent maximum η_1 values containing frequency content $>$1 kHz. Bolded η_2s represent maximum η_2 containing frequency content $<$1kHz. The nasal η_1 is typically much lower than the normal η_1 while the nasal η_2 is higher than the normal η_2.

IMF	η_1:normal	η_2:normal	η_1:nasal4	η_2:nasal4	η_1:nasal7	η_2:nasal7
1	0.0510	1.0000	0.0412	1.0000	0.0149	1.0000
2	0.2714	0.9206	0.0970	0.9274	0.0266	0.9769
3	0.7494	0.5935	0.1487	0.8683	0.0478	0.9661
4	**0.8343**	0.2220	**0.2470**	0.7488	**0.0577**	**0.9393**
5	0.9774	**0.1332**	0.8573	**0.5679**	0.0703	0.9131
6	0.9945	0.0237	0.9818	0.1456	0.6286	0.8835
7	1.0000	0.0051	0.9996	0.0150	0.9995	0.2233
8	1.0000	0.0000	1.0000	0.0001	1.0000	0.0015
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0001
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
11	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 2. Energy metrics: normal voice ’MENRS1’, Level-4 nasal voice ’MENRS4”, and Level-7 nasal voice ’MENRS7’ [10]

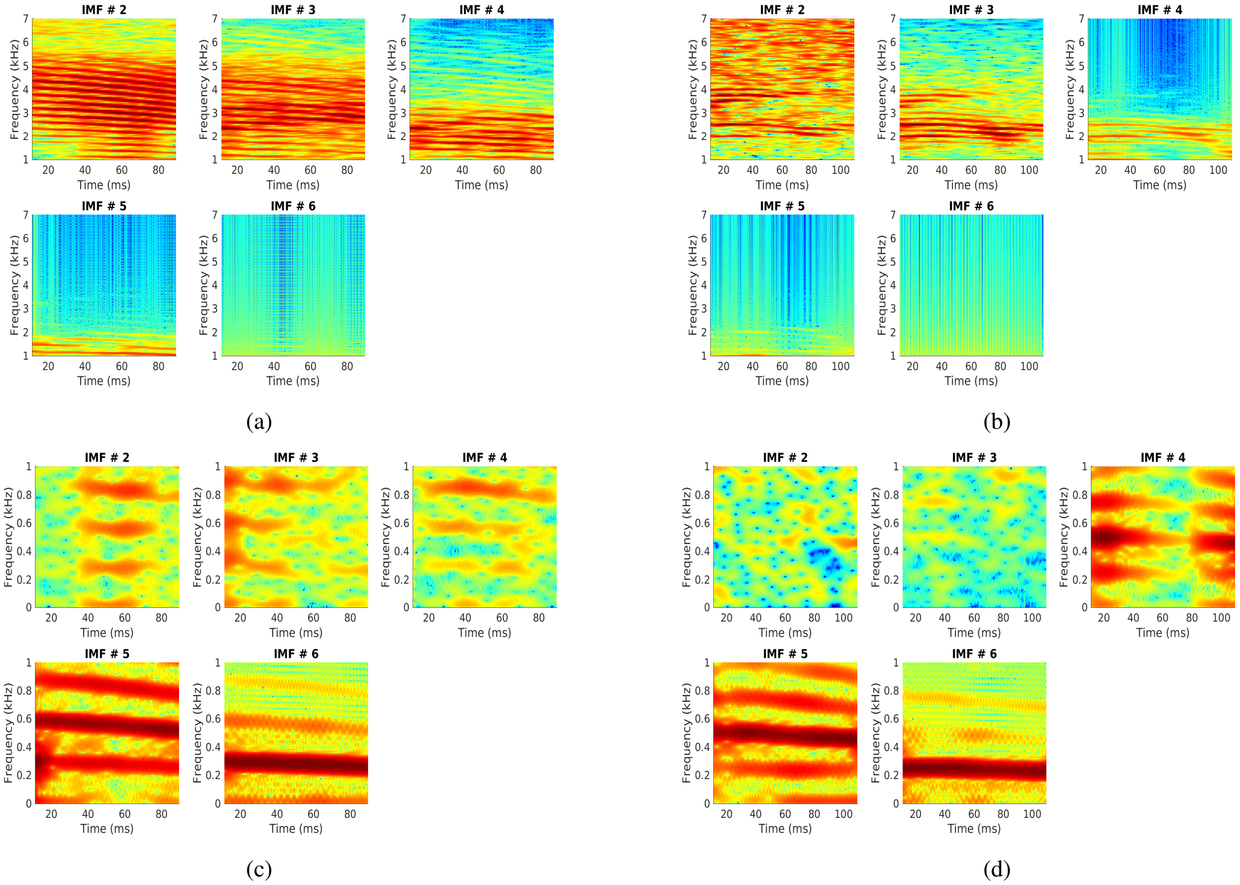
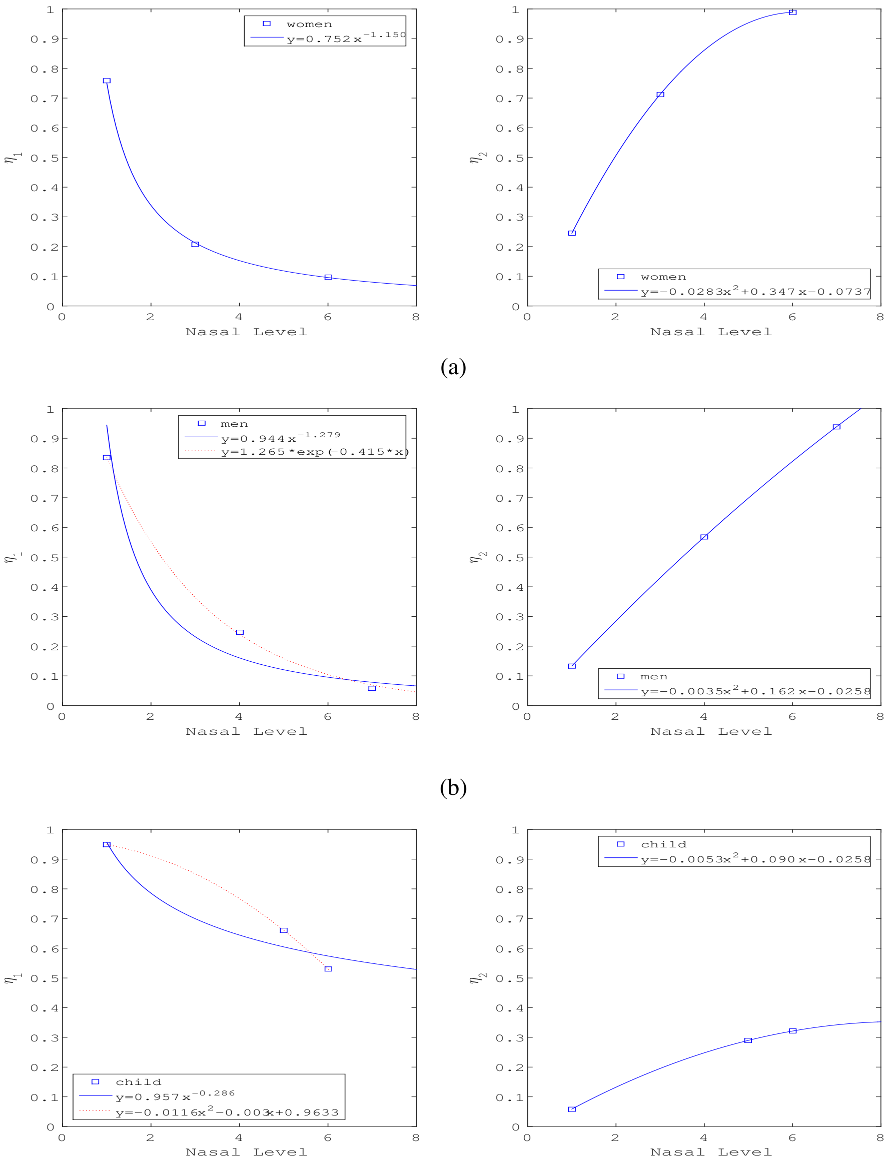


Fig. 2. Spectrograms of IMFs from Figure 1: (a) 1-7 kHz range for normal voice signal, (b) 1-7 kHz range for nasal voice signal, (c) 0-1 kHz range for normal voice signal, and (d) 0-1 kHz range for nasal voice signal. Two hypernasality markers are present: (1) Comparing IMFs #2 and #3 from (a) and (b), a much weaker high-frequency content is observed above 3 kHz for the nasal signal indicating formation of anti-resonances. (2) Comparing IMF #4 from (c) and (d), much stronger low-frequency content is observed in the range 200 Hz - 1 kHz for the nasal signal indicating formation of nasal resonances.

IMF	η_1:normal	η_2:normal	η_1:nasal5	η_2:nasal5	η_1:nasal6	η_2:nasal6
1	0.0062	1.0000	0.0003	1.0000	0.1030	1.0000
2	0.7378	0.9883	0.0028	0.9995	0.3114	0.7636
3	0.9386	0.2476	0.0109	0.9925	0.4445	0.4334
4	**0.9489**	**0.0585**	0.0226	0.9771	**0.5304**	0.3766
5	0.9824	0.0514	**0.6603**	0.9517	0.6185	**0.3214**
6	1.0000	0.0114	0.9998	**0.2900**	0.9370	0.2796
7	1.0000	0.0000	0.9999	0.0002	0.9993	0.0342
8	1.0000	0.0000	1.0000	0.0001	1.0000	0.0007
9	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
10	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
11	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Table 3. Energy metrics: normal voice ’CHILDRS1’, Level-5 nasal voice ’CHILDRS5’, and Level-6 nasal voice ’CHILDRS6’ [10]



(a)

(b)

(c)

Fig. 3. η_1 and η_2 values vs. Nasal level for: (a) Women, (b) Men, (c) Children. Voice samples are from the American CLP-craniofacial database. Nasal levels are determined via perceptual evaluation by clinician, range from 1 to 8, with 1 indicating normal and 8, indicating extreme hypernasality. Note the monotonically decreasing nature of η_1 with increasing hypernasality and monotonically increasing nature of η_2 with increasing hypernasality. Curve fits indicate that the formation of resonances and anti-resonances are fundamentally different and opposite in nature, consistent with the observation that resonances are formed from constructive interference while anti-resonances are formed from destructive interference.

and $m_i[n]$ denotes the i^{th} IMF. The second energy metric measures the energy in the complementary low-frequency bands (less than 1 kHz), $f_s = 44.1$ kHz:

$$\eta_2[k] = \frac{\sum_{i=k}^{n} \tilde{\psi}\left(m_i[n]\right)}{\sum_{i=1}^{n} \tilde{\psi}\left(m_i[n]\right)}. \tag{3}$$

The energy metric η measures the fraction of energy contained in a partial sum of IMFs. For Eq. 1, the summation in the numerator begins at the lowest-numbered IMFs, which contain the high-frequency content, and sums forward. For Eq. 3, the summation in the numerator begins at the highest-numbered IMFs, containing the lowest-frequency content, and sums backward. For both, the partial energy sums are computed as ratios to the median total energy. The TKEO has been used for analysis and detection of hypernasal utterances [4] but in the context of bandpass filtering of utterances for isolating resonances.

3. ANALYSIS AND CLASSIFICATION OF HYPERNASAL VOICE SIGNALS

Figure 1 shows speech samples and the ’relevant’ IMFs of the vowel /i/ for the signals WOMENRS1 and WOMENRS6, where the signal names designate nasal Level-1 (normal) and hypernasal Level-6. Here, a ’relevant’ IMF is one contributing significant energy ($\geq 1\%$) to the total energy spectrum. As observed, the IMFs are oscillatory and are AM–FM signals.

Spectrograms for each IMF are shown in Figure 2. The spectrograms verify that lower-numbered IMFs retain more of the original signal’s high-frequency content while higher-numbered IMFs retain more of the low-frequency content. The spectrograms are used with the energy metrics to establish classification criteria for different hypernasality levels. The energies for each IMF were computed with the TKEO and the energy metrics obtained for the higher and lower frequency bands. Energy metrics (η_1 and η_2 values) for the three women’s voices are shown in Table 1. For completeness, the η values are calculated out to $k = n$. The bolded entries of Table 1 represent the η values that meet the 1 kHz classification criteria and are used to judge the hypernasality level. The frequency content for each IMF is judged from the spectrograms and more precisely determined using the FFT. Table 1 (and the following tables) indicates that: “As nasality increases from Level-1, to Level-4, to Level 6, the energies in the upper frequency bands (η_1) correspondingly decrease from 76%, to 21%, to 9.7%. For the same nasality levels, the energies in the lower frequency bands (η_2) correspondingly increase from 25%, to 71%, to 99%.”

Energy metrics were similarly derived for men and children and are shown in Tables 2 and 3. The dual trends of decreasing η_1 and increasing η_2 as a function of increased hypernasality are observed here as well. Figure 3 shows the η values for all 9 speech signals plotted as a function of nasal level. These trends are consistent across gender and age. In attempting to curve-fit the different η sets, linear regressive power, exponential, and 2nd-order polynomial fits were tested. Intuitively, the energy level should approach zero slope as nasal levels approach a maximum. The best curvatures were convex for the η_1 fits and concave for the η_2 fits. For the η_2 sets, slightly convex behavior was exhibited and therefore all of the η_2 sets were fit to a 2nd-order polynomial.

For the η_1 sets, the power fit works well for women, a exponential fit for the men, and a 2nd-order polynomial fit for children. For a given classification criterion, the η_1 and η_2 metrics delineate between different levels of hypernasal speech. For the vowel /i/, energy metrics for the higher formants (F1-F2) decrease monotonically with increased hypernasality while metrics for the lower formants (F1) increase monotonically with increased hypernasality. The best curve fits for η_1 are either power or exponential and the best fit for η_2 is a 2nd-order polynomial.

4. CONCLUSIONS

In this paper, a novel hybrid EMD/TKEO approach was developed to address the inadequacies of existing approaches for hypernasal speech detection. Traditional techniques for formant analysis such as LPC or one third octave band methods have limitations in analyzing nasal and hypernasal speech signals. Furthermore, these methods were unable to discern between different levels of hypernasality. The hybrid EMD/TKEO approach derived metrics however, were able to detect hypernasality and produce a clear delineation between hypernasality levels when applied to voice samples from the American CLP-craniofacial database.

5. REFERENCES

[1] Alice S-Y. Lee, Valter Ciocca, and Tara L. Whitehill, “Acoustic Correlates of Hypernasality,” *Clinical Linguistics and Phonetics*, Vol. 15, No. 4-5, pp. 259-264, 2003.

[2] Ryuta Kataoka, D. W. Warren, D. J. Zajac, R. Mayo, and R. W. Lutz, “The Relationship Between Spectral Characteristics and Perceived Hypernasality in Children,” *J. Acous. Soc. America*, Vol. 109, No. 5, Part 1, pp. 2181-2189, 2001.

[3] P. Maragos, J.F. Kaiser, and T.F Quatieri, “Energy Separation in Signal Modulations with Applications to Speech Analysis,” *IEEE Trans. Sig. Process.*, Vol. 41, No 10, pp. 3024 - 3051, 1993.

[4] D.A. Cairns, J.H.L Hansen, and J.F. Kaiser, “Recent advances in hypernasal speech detection using the nonlinear Teager energy operator,” *Proc. ICSLP-96*, Vol. 2, pp. 780-783, 1996

[5] M.E. Torres, M.A. Colominas, G. Schlotthauer, and P. Flandrin, “A complete ensemble empirical mode decomposition with adaptive noise,” *Proc. ICASSP-11*, pp. 4144 - 4147, 2011.

[6] R.Deerling and J.F. Kaiser, “The use of a masking signal to improve empirical mode decomposition,” *Proc. ICASSP-05*, Vol. 4, pp. 485 - 488, 2005.

[7] M.A. Colominas, G. Schlotthauer, and M.E. Torres, “Improved complete ensemble EMD: A suitable tool for biomedical signal processing,” *Biomed. Sig. Process. and Control* Vol. 14, pp. 19-29, November 2014

[8] S. Sandoval, P.L. de Leon, and J.M. Liss, “Hilbert spectral analysis of vowels using intrinsic mode functions,” *IEEE ASRU Workshop*, pp. 569 - 575, 2015.

[9] M.M Saidi, P.O. Pietquin, and R. André-Obrecht, “EMD decomposition to discriminate nasal vs. oral vowels in French,” *Proc. SPPRA-10*, pp.128-132, 2010

[10] D. P. Kuehn, P. B. Imrey, L. Tomes, D. L. Jones, M. M. O’Gara , E. J. Seaver, B. E. Smith , D. R. Van Demark, J. M. Wachtel, “Efficacy of Continuous Positive Airway Pressure for Treatment of Hypernasality”, *Cleft Palate Craniofaial Journal* Vol. 39, pp. 267-276, 2002.